様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １５ 日現在

機関番号：１７１０４
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８１１１

研究課題名（和文）最適な作業領域の文法圧縮に基づく索引とストリームデータからの知識発見への応用

研究課題名（英文）String Indexing Based on Space-Optimal Grammar Compression and Its Application to Knowledge Discovery from Stream Data

研究代表者

高畠　嘉将（Takabatake, Yoshimasa）

九州工業大学・大学院情報工学研究院・特任助教

研究者番号：２０８０７０１０

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：繰り返しを多く含む文書データはTBを超えて，今なお増加し続けている．本研究ではそのような増加し続ける巨大ストリームデータを圧縮サイズに比例した作業領域で高速に圧縮可能なデータ圧縮である文法圧縮及びORLBWTを開発した．また，その圧縮データ上で動作する様々な圧縮情報処理技術を開発した．当初の目的である巨大ストリームデータ上のリアルタイムキーワード検索可能な文法圧縮索引の開発には至らなかったが，ORLBWTの構築を実用的に高速化し，そのORLBWT上で動作する高速リアルタイムキーワード検索可能な圧縮索引[Bannai et al. TCS2020]への礎を築いた．

研究成果の学術的意義や社会的意義
開発した文法圧縮やOnline Run-Length BWT (ORLBWT)をTB超のデータをさらに省メモリかつ高速に圧縮可能になった．また，開発したORLBWTを応用したリアルタイムキーワード検索可能な圧縮索引を用いることで巨大なストリームデータから効率的に情報抽出可能となった．また，開発した各種圧縮情報処理技術を応用することで巨大なストリームデータからのリアルタイムの知識発見が可能とすることが期待できる．

研究成果の概要（英文）：Highly repetitive texts exceed TB and are still increasing. In this research, we developed grammar compressions and Online Run-Length BWTs (ORLBWTs), which can compress such large streaming data at high speed in compressed space. Furthermore, we developed various information processes on the compressed data. Although we could not develop a grammar-based compressed index supporting real-time keyword searches on large streaming data, we significantly improved the construction time of ORLBWTs and our ORLBWTs resulted in the development of an ORLBWT-based compressed index supporting real-time searches on large streaming data [Bannai et al. TCS2020].

研究分野：文字列のデータ圧縮とその圧縮データ上での情報検索

キーワード：データ圧縮　圧縮索引　圧縮情報処理　文法圧縮　BWT

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

Wikipedia の文書や GitHub のソースコードなどの編集履歴を保存した世代管理システム，1000 人ゲノム計画などの同じ生物のゲノムを保存したデータベースでは同じ部分文字列が異なる位置に繰り返し出現する文書データが TB を超えて，今なお増加し続けている．これらの増加し続けるデータであるストリームデータの増加に対応するために①高速かつ省メモリに小さく圧縮する圧縮法が必要である．さらに，これらのデータから有用な文書やソースコードの検索・抽出やゲノムの解析といった再利用を行うために，これらの再利用を高速に実行するための②情報検索技術も必要である．

①に関して，一般的なデータ圧縮では，ストリームデータのようにデータの末尾への追加の必要な場合は追加のたびに解凍して，最初から圧縮し直す必要があるため，TB 超のデータでは時間がかかりすぎてしまうか，作業領域（メモリ使用量）が入力データ以上かかってしまうかのどちらかである．そこで申請者は圧縮データに新たなデータを高速に末尾追加可能で圧縮サイズに比例した作業領域で圧縮可能な文法圧縮と呼ばれるデータ圧縮の SOLCA を開発し，理論と実用の両面から TB 超のストリームデータを省メモリかつ高速に小さく圧縮することを可能としている．

②に関してはキーワード検索を行うための接尾辞配列や接尾辞木などの通常の索引では入力データサイズの作業領域が必要なため，文法圧縮，RLBWT や LZ77 などの圧縮データ上での高速キーワード検索技術である圧縮索引が提案されている．しかし，通常の圧縮索引は新たなデータに対して再構築が必要なため，TB 超のストリームデータに対応できない．そこで，申請者は圧縮サイズに比例した作業領域で新たなデータを高速に末尾追加可能な文法圧縮を応用した圧縮索引の理論 OESP-index [1]を開発したが，その実際の構築速度は遅すぎるし，作業領域もまだ多く，TB 超のストリームデータからリアルタイムの検索をするのは難しい．

したがって，TB 超のストリームデータに対して，①と②を同時に可能であり，かつ②やそれを応用した知識発見をリアルタイムに行えるようにすることは現在の圧縮索引技術では計算時間もしくは作業領域のどちらかの観点から困難である．

２．研究の目的

そこで，本研究では SOLCA の圧縮速度と圧縮サイズ，世界最小の作業領域を保ったまま，リアルタイムのキーワード検索可能な圧縮索引へ拡張することで，OESP-index の構築速度，作業領域，検索速度の理論と実用の両面から改善し，TB 超のストリームデータを省メモリかつ高速に圧縮しながら，キーワード検索をリアルタイムに可能であることを目指す．さらに，そのキーワード検索を剽窃検出のための曖昧検索や頻出キーワード検出に応用することで，Wikipedia や GitHub からの有用なデータの抽出と剽窃検出やゲノムの解析といった TB クラスのストリームデータからの知識発見をリアルタイムに可能とする技術の創出を目的とする．

３．研究の方法

データ圧縮およびその圧縮データ上で圧縮データサイズに比例した作業領域で動作する圧縮情報処理の各種アルゴリズムを設計及び実装し，計算機実験を行い，各種アルゴリズムの評価を行った．

４．研究成果

本研究の主な成果は以下の通りである．

(1) 文法圧縮の一種である RePair は文法圧縮の中で世界最小に圧縮可能である一方で，メモリ使用量が入力データの 10 倍以上必要である．この理論研究では，その RePair のメモリ使

用量を極限まで減らした RePair の世界初の In-place アルゴリズムを開発した．In-place アルゴリズムとは入力データサイズに定数を足したメモリ使用量で動作するアルゴリズムである．

(2) RePair のメモリ使用量を実用的に減らしつつ，高速に計算するアルゴリズムを開発した．この方法ではあらかじめ省メモリかつ高速に動作する圧縮法で圧縮しておき，その圧縮データから RePair の圧縮データに変換する再圧縮という技術を用いることで 既存の RePair に対して，約 10%のメモリ使用量でかつ 30 倍から数百倍の高速化を果たした．

(3) 他の文法圧縮から RePair への再圧縮を圧縮データサイズのメモリ使用量でかつ圧縮データサイズに比例した計算時間で可能な世界初の理論を構築した．SOLCA などの圧縮データサイズに比例したメモリ使用量しか使わない文法圧縮からこの RePair への再圧縮を行うことで，圧縮データサイズに比例したメモリ使用量のみで RePair を計算可能にした．

(4) 文法圧縮の一種である LZ78 を基盤とした LZD は実用的に小さく圧縮可能である一方で，その素朴な計算方法では理論的に遅い計算時間の下界[4]が知られている．その計算時間の下界を避けることが可能な LZ-ABT を開発した．

(5) 申請者が開発した ESP-index-I[2]は長いキーワード文字列を高速検索可能な文法圧縮索引である一方で，短いキーワード文字列に対して，他の圧縮索引と比べて低速であった．そのため，短いキーワード文字列のための圧縮データサイズの新たなデータ構造を追加することで，2 から 10,000 倍の高速化を果たした．

(6) 圧縮データサイズのメモリ使用量で動作し，かつ高速に末尾追加可能な Online Run-Length BWT (ORLBWT)と呼ばれるデータ圧縮を実用的に高速化した．実験では最大 60 倍の高速化を達成した．なおこの高速化技術が元となり，圧縮データサイズのメモリ使用量で構築可能でかつ高速な末尾追加も可能な圧縮索引[3]が開発されている．

(7) 文法圧縮されたデータから全てのデータを復元することなく，任意の一部分のデータを高速に復元する技術を既存の手法とほぼ同等のメモリ使用量を保ちつつ，約 3.6 倍の高速化を果たした．

(8) 世界初の移動付き編集距離の秘匿計算技術を開発した．移動付き編集距離は 2 つの文字列間の類似度を測る指標の一つである．移動付き編集距離の秘匿計算とは 2 人の人間が持つ文字列間の移動付き編集距離をお互いが持っている文字列がバレないように計算可能な技術である．

（参考文献）

[1] Yoshimasa Takabatake, Yasuo Tabei and Hiroshi Sakamoto: Online Self-Indexed Grammar Compression. In Proceedings of the 22nd International Symposium on String Processing and Information Retrieval (SPIRE), LNCS9309, pages 258-269, 2015.

[2] Yoshimasa Takabatake, Yasuo Tabei and Hiroshi Sakamoto: Improved ESP-index: A Practical Self-index for Highly Repetitive Texts. In Proceedings of the 13th International Symposium on Experimental Algorithms (SEA), LNCS8504, pages 338-350, 2014.

[3] Hideo Bannai, Travis Gagie and Tomohiro I: Refining the r-index, Theoretical Computer Science, 812:96-108, 2020.

[4] Golnaz Badkobeh, Travis Gagie, Shunsuke Inenaga, Tomasz Kociumaka, Dmitry Kosolobov and Simon Puglisi, On Two LZ78-style Grammars: Compression Bounds and Compressed-Space Computation, In Proceedings of the 24th International Symposium on String Processing and Information Retrieval (SPIRE), LNCS 10508, pp. 51-67, 2017.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Koppl Dominik、I Tomohiro、Furuya Isamu、Takabatake Yoshimasa、Sakai Kensuke、Goto Keisuke	４．巻 14
２．論文標題 Re-Pair in Small Space	５．発行年 2020年
３．雑誌名 Algorithms	６．最初と最後の頁 5～5
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/a14010005	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Nishimoto Takaaki、Takabatake Yoshimasa、Tabei Yasuo	４．巻 273
２．論文標題 A compressed dynamic self-index for highly repetitive text collections	５．発行年 2020年
３．雑誌名 Information and Computation	６．最初と最後の頁 104518～104518
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.ic.2020.104518	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Ohno Tatsuya、Sakai Kensuke、Takabatake Yoshimasa、I Tomohiro、Sakamoto Hiroshi	４．巻 52-53
２．論文標題 A faster implementation of online RLBWT and its application to LZ77 parsing	５．発行年 2018年
３．雑誌名 Journal of Discrete Algorithms	６．最初と最後の頁 18～28
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.jda.2018.11.002	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　9件）

１．発表者名 Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Louisa Seelbach Benkner, Yoshimasa Takabatake
２．発表標題 Practical Random Access to SLP-Compressed Texts
３．学会等名 The 27th International Symposium on String Processing and Information Retrieval（国際学会）
４．発表年 2020年

１．発表者名
Dominik Koppl , Tomohiro I, Isamu Furuya, Yoshimasa Takabatake, Kensuke Sakai, Keisuke Goto

２．発表標題
Re-Pair in Small Space

３．学会等名
Prague Stringology Conference（国際学会）

４．発表年
2020年

１．発表者名
Yohei Yoshimoto, Masaharu Kataoka, Yoshimasa Takabatake,Tomohiro I, Kilho Shin, Hiroshi Sakamoto

２．発表標題
Faster Privacy-Preserving Computation of Edit Distance with Moves

３．学会等名
The 14th International Workshop on Algorithms and Computation（国際学会）

４．発表年
2020年

１．発表者名
Dominik Koppl , Tomohiro I, Isamu Furuya, Yoshimasa Takabatake, Kensuke Sakai, Keisuke Goto

２．発表標題
Re-Pair in Small Space

３．学会等名
Data Compression Conference（国際学会）

４．発表年
2020年

１．発表者名
Travis Gagie, Tomohiro I, Giovanni Manzini, Gonzalo Navarro, Hiroshi Sakamoto, Yoshimasa Takabatake

２．発表標題
Rpair: Rescaling RePair with Rsync

３．学会等名
The 26th International Symposium on String Processing and Information Retrieval（国際学会）

４．発表年
2019年

１．発表者名
Kensuke Sakai、Tatsuya Ohno、Keisuke Goto、Yoshimasa Takabatake、Tomohiro I、 Hiroshi Sakamoto

２．発表標題
RePair in Compressed Space and Time

３．学会等名
Data Compression Conference（国際学会）

４．発表年
2019年

１．発表者名
Shunta Nakagawa、Tokio Sakamoto、Yoshimasa Takabatake、Tomohiro I、Kilho Shin、Hiroshi Sakamoto

２．発表標題
Privacy-Preserving String Edit Distance with Moves

３．学会等名
The 11th International Conference on Similarity Search and Applications（国際学会）

４．発表年
2018年

１．発表者名
Tatsuya Ohno、Keisuke Goto、Yoshimasa Takabatake、Tomohiro I、Hiroshi Sakamoto

２．発表標題
LZ-ABT: A Practical Algorithm for α-Balanced Grammar Compression

３．学会等名
The 29th International Workshop on Combinatorial Algorithms（国際学会）

４．発表年
2018年

１．発表者名
Reimi Tanaka、Yoshimasa Takabatake、Tomohiro I、Hiroshi Sakamoto

２．発表標題
Improved Grammar Compression in Constant Space

３．学会等名
The 14th International Conference on Grammatical Inference（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
カナダ	Dalhousie University			
イタリア	University of Piemonte Orientale			
チリ	University of Chile			
ドイツ	University of Siegen			